Ronak Haresh Chhatbar

+1 716-507-2419 | ronakchhatbar@gmail.com | LinkedIn | GitHub | alphapibeta.com | Buffalo, New York

Professional Summary

Production AI systems engineer working across the full inference stack — **GPU optimization, real-time computer vision, and agentic LLM orchestration** — from edge hardware to enterprise cloud. Architected core modules of Centific's **Verity** platform, cutting facility monitoring costs **33%** and incident response time **67%**. Five years across Jetson edge, on-premises NVIDIA GPU infrastructure, and cloud. Designs and operates a self-hosted multi-model inference stack serving live production traffic.

Experience

- **Centific** — Remote
  AI Engineer — Verity Multimodal Agentic Platform — *Dec 2024 – Present*
  - **Owned the inference-to-orchestration stack** on **Verity** — Centific's NVIDIA-accelerated enterprise vision platform connecting live video, sensor networks, and operational systems — cutting **24/7 facility monitoring costs 33%** by replacing continuous human surveillance with autonomous AI-driven incident detection across enterprise deployments.
  - **Engineered the real-time CV pipeline** at the core of Verity: GPU-accelerated ingestion of **50+ concurrent camera streams**, frame-level object detection, and chained LLM inference for contextual incident classification — delivering **sub-second detection latency** end-to-end at production scale.
  - **Cut incident response time from 15 to 5 minutes** by designing a **LangGraph stateful agent graph** with conditional routing — specialized sub-agents invoke live sensor queries, facility maintenance APIs, and compliance report generators, coordinated through a shared state machine driven by voice and text input.
  - **Reduced inference infrastructure costs 20%** by engineering Verity's model serving layer with **TensorRT-LLM** — implementing concurrent multi-model execution with GPU memory partitioning across co-deployed vision and language models to sustain real-time SLAs under production load.
  - **Designed the operational data layer** powering Verity's situational awareness: a Python/SQL ingestion pipeline unifying camera event streams, IoT sensor telemetry, and maintenance APIs into a real-time dashboard — eliminating cross-system manual queries during incident triage.

- **Tensorgo Technologies** — Hyderabad, India
  Computer Vision Engineer — *Sep 2020 – Aug 2022*
  - **Improved model throughput 25% and inference performance 40%** by going deep on GPU-level optimization — applied **TensorRT and DeepStream** mixed-precision tuning and GPU memory profiling to eye-gaze and emotion recognition models, sustaining **30–40 FPS** on Jetson NX and Nano under production memory constraints.
  - **Achieved 8% accuracy improvement** in contactless heart rate estimation (rPPG from video) by training on **20,000+ images** across BP4D+, UBFC-1, and UBFC-2 — explicitly modeling demographic variation in skin tone and lighting for inclusive biometric deployment.
  - **Increased meeting analytics accuracy 16%** by integrating real-time speaker segmentation into the emYt+ compliance platform via an ASR pipeline — extending the CV stack into multimodal audio+video analysis serving enterprise Zoom and Webex integrations.
  - **Reduced manual intervention 60%** and deployment time **30%** by building automated training, evaluation, and Jetson edge deployment pipelines — enabling continuous delivery of updated CV models to production hardware without manual optimization steps.

- **Wavelabs Technologies** — Hyderabad, India
  Machine Learning Engineer — *May 2019 – Aug 2020*
  - **Delivered sub-2-second threat identification at 30–40 FPS** on resource-constrained Jetson Nano by building a real-time weapon detection system integrated with iOS/Android mobile apps — overcoming embedded hardware limits through model optimization and data scale (**150,000+ manually labeled and augmented images**).
  - **Generated 15% revenue increase** by developing dynamic pricing algorithms across **10,000+ monthly transactions** for financial services clients.
  - **Reduced resource utilization 35%** and deployment time **20%** by containerizing model training and serving pipelines with **Docker** and **AWS SageMaker** — establishing reproducible workflows across development and production environments.

Projects & Research

- **Self-Hosted Multi-Model AI Stack**
  Production Agentic Inference Infrastructure — Full Stack: GPU · CV · Agents · Voice
  - Architected and operate a **live production AI stack**: **SGLang** serving a 20B LLM at 63 tok/s (2×GPU tensor-parallel) alongside **Gemma 4** on RTX 2060 at 94 tok/s for vision+reasoning, and **NanoOWL** (OWL-ViT TRT engine on Jetson Orin Nano, ∼300 ms) for open-vocabulary object detection — three nodes, eight services, wired local network.
  - Built a custom Python **agentic loop** with dynamic routing across models, streaming SSE, tool-calling via **Model Context Protocol (MCP)**, and **RAG** over Qdrant with bge-base-en-v1.5 embeddings and a cross-encoder reranker — full retrieval round-trip under 150 ms on Jetson NVMe.
  - Engineered an **end-to-end voice pipeline** (faster-whisper ASR → LLM → Kokoro GPU TTS) delivering first audio in **under 4 seconds**; routed through a LiteLLM proxy with LangFuse distributed tracing across all nodes.

- **Spatial AI & Robotics Lab** University at Buffalo
  Graduate Research Assistant *Dr. Chen Wang — May 2023 – Dec 2024*
  - **Increased visual odometry inference efficiency 33%** by developing C++ plugins integrating a custom visual navigation optimizer with a TensorRT backend, enabling real-time pose estimation on resource-constrained platforms.
  - Led backend development of robotranking.org, a robotics benchmarking platform adopted as a reference resource by the international robotics research community.
  - Improved optical flow estimation accuracy and real-time throughput by applying specialized interpolation algorithms to dense frame sequences in dynamic environments.

- **GPU Computing Research Portfolio**
  Advanced CUDA Development
  - **Hessian Matrix Inversion**: Achieved **526× GPU speedup** over CPU baseline via LU decomposition with cuSOLVER and Python bindings.
  - **CUDA Performance Profiler**: Automated Nsight Compute metrics collection with an interactive Streamlit dashboard for systematic kernel optimization.
  - **Convolution Optimization Analysis**: Published analysis across **18 optimization metrics** with mathematical foundations and Nsight Compute profiling visualization.

Education

- **University at Buffalo, The State University of New York** Buffalo, NY
  *M.S. Computer Science — GPA: **3.4/4.0*** *Aug 2022 – Jan 2024*
  Courses: Operating Systems, Analysis of Algorithms, Biometrics Image Analysis, Reinforcement Learning, Computer Vision

- **Jawaharlal Nehru Technological University Hyderabad** Hyderabad, India
  *B.E. Computer Science — GPA: **3.6/4.0*** *Aug 2015 – May 2019*
  Courses: Machine Learning, Cloud Computing, Data Structures & Algorithms, Computer Networks, Probability & Statistics

Skills

- **GPU Computing & Inference Optimization**: CUDA, TensorRT, TensorRT-LLM, NVIDIA DeepStream, Nsight Compute, Mixed Precision, Triton Inference Server, OpenMP
- **Agentic AI & LLM Systems**: LangGraph, LangChain, Model Context Protocol (MCP), Multi-Agent Orchestration, RAG Pipelines, Qdrant, SGLang, vLLM, LiteLLM, LangFuse
- **Computer Vision & Machine Learning**: PyTorch, TensorFlow, OpenCV, ONNX, Real-Time Video Processing, Object Detection, ASR Integration, Biometric Systems
- **Infrastructure & Languages**: Python, C/C++, CUDA, SQL, Docker, Kubernetes, AWS, Jetson (Orin/NX/Nano), Linux, PostgreSQL, Vector Databases, Kafka